Opinion

Open Access CrossMark

Next-generation sequencing or the dilemma of large-scale data analysis: opportunities, insights, and challenges to translational, preventive and personalized medicine

Abstract

Over the past years, the advent of Next-Generation Sequencing (NGS) technology, also known as high-throughput sequencing (HTS), has represented an immense hope for all of us. Indeed, NGS has (r)evolutionized the fields of molecular biology, genetics and genomics, enabling cost-effective and quick generation of DNA and RNA sequence data with exquisite accuracy and resolution for possible translational, preventive and personalized medicine. Nevertheless, in spite of tremendous advancements to broaden NGS applications from research to clinic, NGS still presents enormous challenges in terms of data storage, processing, quality control management and interpretation, which slow the translation from the bench-top to the bed-side. In this expert-opinion article, I first summarize the main doubts about NGS technology according to my experiences in the field, which actually could open-up new opportunities for innovative research and development. I further highlight the general technological and methodological characteristics of NGS as well as the recent advances and challenges in terms of clinical investigations and applications toward the development of theranostics. Eventually, I briefly question the relevance of integrating NGS with other platforms such as next-generation proteomics (NGP) to optimize the prognosis, diagnosis and therapeutic options.

Keywords: next-generation sequencing, dna-seq, rna-seq, next-generation proteomics, high-throughput screening, theranostics, molecular diagnosis, genetic-based prognosis, personalized medicine, translational medicine, innovative technologies

Volume 1 Issue 1 - 2014

Farid Menaa
Department of Oncology, Stem cells and Nanomedicine, Fluorotronics Inc., USA

Correspondence: Farid Menaa, Department of Oncology, Stem cells and Nanomedicine, Fluorotronics Inc., 2453 Cades Way, Vista, CA 92081, USA, Email dr.fmenaa@gmail.com

Received: May 01, 2014 | **Published:** May 24, 2014

Abbreviations: NGS, next-generation sequencing; HTS, high-throughput sequencing; NGP, next-generation proteomics; SNPs, single-nucleotide polymorphisms, CNVs, copy number variations

Next-generation sequencing: between hopes and doubts for new opportunities!

Demand for faster DNA sequencing data than Sanger sequencing methods, which was used to sequence the first human genome, has led to NGS technologies.[1,2] NGS allows genomes analyses, including those representing complex disease states such as cancers[3–6] or hematological disorders.[7] In many cases, NGS, and so integrative genomics, can solve unmet clinical needs in the diagnosis, prediction of prognosis, monitoring the status of the disease and personalized treatment decision.[8–10] In fact, NGS is (R) evolutionizing the genetic field, bringing both hopes and doubts for many people including scientists, researchers, health practitioners, and patients. Thereby, high volume of data obtained from NGS nucleic acids sequencing instruments (DNA-seq, RNA-seq) leads to several new kinds of experiments, new questions amenable to study and new challenges necessary to efficient translational medicine.[11] Indeed, how do we get from a collection of several million short sequences of bases to genome-scale results? Can we accurately translate these data from bench-top to bed-side? Can we reliably use them for preventive and personalized medicine? How important is to consider ethnical genetic diversity for genomic data interpretation? How long shall we keep the data and the samples of a given patient, considering possible spontaneous or induced genomic alterations? What are the risks for the patient to have his whole nucleic acid sequenced? What are the ethical challenges? How the static genomic information can be interpreted in the dynamic molecular world? What genetic information can be considered as a driver (disease) event? Shall we focus on homogeneous cellular/tissues subgroups or shall we use heterogeneous biological material (e.g. pool of cells, whole tissue) to identify genetic aberrations as therapeutic targets? What is the best sequencing platform and methodology to use? What is the cost of performing an accurate NGS of our genome? Are meta-analyses from both epidemiological and genomic studies required? What reference(s) one shall use to interpret NGS data? What is the confidence level of getting true information? How to fill all the gaps to obtain more confident information? Can we rely only on genome sequencing data? Shouldn´t we consider the combination of DNA-NGS, RNA-NGS and Next-Generation Proteomics (NGP) data analyses to obtain a better comprehensive view of the molecular dynamics and develop more accurate theranostics?

DNA-SEQ and RNA-SEQ: basic framework, applications and challenges to translational, preventive and personalized medicine

While much discussion focuses on rapidly sequencing human genomes at a low cost, the grail of personalized genomics, a vast amount of research must be performed at the systems level to fully understand the relationship between biochemical processes in a cell and how the instructions for the processes are encoded in the genome. Systems biology and a plethora of "omics" have emerged to measure multiple aspects of cell biology as DNA is transcribed into RNA and RNA translated into protein and proteins interact with molecules to carry out biochemistry. DNA NGS is being used to perform quantitative assays where DNA sequences are used as highly informative data points. In these assays, large datasets of sequence reads are collected in a massively parallel format. Reads are aligned to reference data to obtain quantitative information by tabulating the frequency, positional information, and variation from the reads in the alignments. Data tables from samples that differ by experimental treatment, environment, or in populations, are compared in different ways to make discoveries (e.g. mutations, single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), methylation and/or acetylation sites) and draw experimental conclusions.

In practice, NGS data analysis process involves three stages. At the first stage, i.e. *primary data analysis*, image data are converted to sequence data. The sequence data (reads) can be in familiar "ACTG" sequence space or less familiar color space (SOLiD) or flow space. Primary analyses also provide quality values for each base that are used in subsequent phases of analysis, much like Phred quality values were used in Sanger sequencing. In the middle stage, i.e. *secondary data analysis*, datasets are created. Sequences from the primary data are aligned to reference data (e.g. complete genomes, subsets of genomic data like expressed genes, individual chromosomes) to create application-specific data sets for each sample. Presently, there is a large and growing list of alignment programs that can be used for secondary data analysis. In the final stage, i.e. *tertiary data analysis*, the data sets are compared to create experiment-specific results. This phase may involve a simple activity, like viewing a dataset in genome browser and using the frequency of tags to identify promoter sites or patterns of variation. Other experiments, like digital gene expression, include tertiary analyses where datasets are compared to each other, as it is done with microarray data. These kinds of analyses are the most complex: expression measurements need to be normalized between datasets and statistical comparisons need to be made to assess differences. Currently, the software for the primary analyses is provided by the instrument manufacturers and handled within the instrument itself, and when it comes to the tertiary analyses, many good tools already exist. However, between the primary and tertiary analyses lies a gap, but emerging studies reported advanced strategies and showed that NGS brings more robustness, resolution and inter-lab portability than microarray platforms.[12] Thereby, robust mutation detection can be obtained by NGS assays if the data can be processed in a way (e.g. using artificial amplicon data sets) that does not lead to large genomic alterations landing in the unmapped data (i.e. trash).[13]

In RNA-Seq, the process of determining relative gene expression means that sequence data from multiple samples must go through the entire process of primary, secondary, and tertiary analysis. To do this work, researchers must puzzle through a diverse collection of early version algorithms that are combined into complicated workflows with steps that produce complicated file formats. Command line tools such as MAQ, SOAP, MapReads, and BWA, have specialized requirements for formatted input and output and leave researchers with large data files that still require additional processing and formatting for tertiary analyses. Moreover, once reads are aligned, datasets need to be visualized and further refined for additional comparative analysis. Solutions to these challenges that close the gaps between primary, secondary, and tertiary analyses by showing results from a complete workflow system that includes data collection, processing and analysis for expression analysis are being developed. NGS is an attractive option for analyzing a transcriptome because the vast numbers of reads that can be obtained along with their sequences provide a highly sensitive way to evaluate the RNA population inside of a cell.[14] In addition to rRNA, tRNA, and mRNA, assays are also measuring non-coding RNA and multiple classes of small RNAs (e.g. miRNA), but not without risks of biases.[15–17] As one obtains deeper information, largely through NGS, one learns that even mRNA is more complicated than previously thought. New reports indicate that 92-97% of human genes undergo alternative splicing.[18,19] A common goal for these assays is to map the structure of genes in terms of their start sites, 5' and 3' ends, exons, splice junctions, polyA sites, and alternative forms, and quantify the relative abundance of different molecules under different conditions or developmental stages. When considered in an NGS context, transcriptome analysis breaks into categories of experiments defined by different procedures and analysis paths. Despite the widespread utilization of NGS, a major bottleneck in the implementation and capitalization of this technology therefore remains in the data processing steps.[13] Further, the brisk evolution of sequencing technologies has flooded the market with commercially available sequencing platforms, whose unique chemistries and diverse applications stand as another obstacle restricting the potential of NGS for clinical applications.[2] Importantly, large consortium-based sequencing studies (e.g. candidate gene studies, genome-wide association studies, and whole-genome admixture-based approaches that account for ancestral genetic structure, complex haplotypes, gene-gene interactions, and rare variants to detect and replicate novel pharmacogenetic loci) are using next-generation whole-genome sequencing to provide a diverse genome map of different admixed populations, which can be used for future pharmacogenetic studies.[10] Therefore, it is time to work together more closely and move forward with awareness and holistic knowledge of the NGS capabilities and applications to the clinical realm. Interestingly, strategies for addressing the challenges of implementing genomic medicine in the clinical setting with more accuracy are emerging through best practices for integrating genomic findings into the electronic health record (e.g. eMERGE network).[20]

Technological integration of next-generation sequencing and next-generation proteomics: let´s think a step forward!

Eventually, in complement to DNA-seq and RNA-seq, and because the causes of most disorders are multi-factorial, another system-level approach to be considered, for a more comprehensive understanding of human biological complexity, is to integrate a view of proteome dynamics, possibly using MS-based proteomics.[21] It seems that it is never too early to think about the biggest challenges for more accurate, more efficient and safer molecular-based medicine!..

Conclusion

NGS constitutes a major breakthrough in genomic research. Despite the advantages of NGS platforms related to the HTS rate and cost-effectiveness, the assembly of the reads produced by the current next-generation sequencers still remains a major challenge to faster translation to clinic. Maybe a layered architecture approach for constructing a general assembler that can handle the sequences generated by the different available sequencing platforms is a potent solution? Also, shall we already think to implement or integrate NGS with other technologies for a common work at the different system-levels?

Acknowledgements

None.

Conflict of interest

Author declares that there is no conflict of interest.

References

1. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135–1145.
2. Rizzo JM, Buck MJ. Key principles and clinical applications of “next-generation” DNA sequencing. *Cancer Prev Res (Phila)*. 2012;5(7):887–900.
3. Wu K, Huang RS, House L, et al. Next-generation sequencing for lung cancer. *Future Oncol*. 2013;9(9):1323–1336.
4. Lee EJ, Luo J, Wilson JM, et al. Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer Lett*. 2013;340(2):171–178.
5. Yamamoto H, Watanabe Y, Maehata T, et al. An updated review of gastric cancer in the next-generation sequencing era: Insights from bench to bedside and vice versa. *World J Gastroenterol*. 2014;20(14):3927–3937.
6. Burrell RA, Swanton C. The evolution of the unstable cancer genome. *Curr Opin Genet Dev*. 2014;24:61–67.
7. Johnsen JM, Nickerson DA, Reiner AP. Massively parallel sequencing: the new frontier of hematologic genomics. *Blood*. 2013;122(19):3268–3275.
8. Lee SH, Sim SH, Kim JY, et al. Application of cancer genomics to solve unmet clinical needs. *Genomics Inform*. 2013;11(4):174–179.
9. Natrajan R, Wilkerson P. From integrative genomics to therapeutic targets. *Cancer Res*. 2013;73(12):3483–3488.
10. Ortega VE, Meyers DA. Pharmacogenetics: implications of race and ethnicity on defining genetic profiles for personalized medicine. *J Allergy Clin Immunol*. 2014;133(1):16–26.
11. Kamalakaran S, Varadan V, Janevski A, et al. Translating next generation sequencing to practice: opportunities and necessary steps. *Mol Oncol*. 2013;7(4):743–755.
12. Hoen PA, Ariyurek Y, Thygesen HH, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res*. 2008;36(21):e141.
13. Daber R, Sukhadia S, Morrissette JJ. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. *Cancer Genet*. 2013;206(12):441–448.
14. Cloonan N, Forrest AR, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*. 2008;5(7):613–619.
15. Zamore PD, Haley B. Ribo-gnome: the big world of small RNAs. *Science*. 2005;309(5740):1519–1524.
16. Anglicheau D, Muthukumar T, Suthanthiran M. MicroRNAs: small RNAs with big effects. *Transplantation*. 2010;90(2):105–112.
17. Raabe CA, Tang TH, Brosius J, et al. Biases in small RNA deep sequencing data. *Nucleic Acids Res*. 2014;42(3):1414–1426.
18. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470–476.
19. Castle JC, Zhang C, Shah JK, et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet*. 2008;40(12):1416–1425.
20. Kullo IJ, Haddad R, Prows CA, et al. Return of results in the genomic medicine projects of the eMERGE network. *Front Genet*. 2014;5:50.
21. Altelaar AF, Munoz J, Heck AJ. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet*. 2013;14(1):35–48.

Citation: Menaa F. Next-generation sequencing or the dilemma of large-scale data analysis: opportunities, insights, and challenges to translational, preventive and personalized medicine. *J Investig Genomics*. 2014;1(1):27–29. DOI: 10.15406/jig.2014.01.00005